MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-822

An Efficient Pancreatic Cyst Identification Methodology Using Natural Language Processing

Saeed Mehrabi[a], C. Max Schmidt[b], Joshua A. Waters[b], Chris Beesley[c], Anand Krishnan[a], Joe Kesterson[c], Paul Dexter[c], Mohammed A. Al-Haddad[d], William M. Tierney[c,d], Mathew Palakal[a]

[a] *School of Informatics, Indiana University, Indianapolis, IN, USA*
[b] *Department of Surgery, Indiana University, Indianapolis, IN, USA*
[c] *Regenstrief Institute, Indianapolis, IN, USA*
[d] *Department of Medicine, Indiana University, Indianapolis, IN, USA*

Abstract

Pancreatic cancer is one of the deadliest cancers, mostly diagnosed at late stages. Patients with pancreatic cysts are at higher risk of developing cancer and their surveillance can help to diagnose the disease in earlier stages. In this retrospective study we collected a corpus of 1064 records from 44 patients at Indiana University Hospital from 1990 to 2012. A Natural Language Processing (NLP) system was developed and used to identify patients with pancreatic cysts. NegEx algorithm was used initially to identify the negation status of concepts that resulted in precision and recall of 98.9% and 89% respectively. Stanford Dependency parser (SDP) was then used to improve the NegEx performance resulting in precision of 98.9% and recall of 95.7%. Features related to pancreatic cysts were also extracted from patient medical records using regex and NegEx algorithm with 98.5% precision and 97.43% recall. SDP improved the NegEx algorithm by increasing the recall to 98.12%.

Keywords:

Pancreatic cyst, Natural language processing, Unstructured Information Management Architecture, negation, dependency parser.

Introduction

Pancreatic cancer is the fourth leading cause of cancer death in the US [1] with estimated 44,920 new cases and 37,390 deaths in 2012. It is the deadliest cancer with the overall five-year survival rate of 4.4% and little improvement of survival over the past 30 years [2]. Despite advances in surgery, radiotherapy and chemotherapy more than 90% of patients with pancreatic cancer die of chemo insensitive disease. Although there is no cure for pancreatic cancer, it can be prevented in patients with heredity predisposition to pancreatic cancer and patients with pre-cancerous pancreatic cysts [3]. Pancreatic cysts are present in an estimated 2-3% of the adult population. Pancreatic cysts have degrees of malignancy based upon their type and degree of dysplasia. There are various types of pancreatic cysts which include pseudocysts, serous cystic neoplasms (SCN), mucinous cystic neoplasms (MCN) and intraductal papillary mucinous neoplasms (IPMN). Only MCN and IPMN may progress to invasive pancreatic adenocarcinoma [4]. Accurate identification, surveillance and treatment of pancreatic cysts represents an opportunity to prevent pancreatic cancer. Much information about pancreatic cysts can be found in free text format in various narrative medical reports including pathology, cytopathology, radiology (MRI, CT, EUS) and physician's clinical reports. Once a cyst patient has been identified, patients' symptoms (pancreatitis, diabetes, jaundice, etc.), radiology findings (mural nodules, main pancreatic duct dilation, mass, etc.), pathology findings (Atypia), lab findings (Ha1c, Carcinogenic Antigen 19-9, etc.), and molecular data (KRas mutation) can be extracted which aid in determining a patient's risk of cyst progression to cancer. In a series of previous works, we studied the treatment management of pancreatic cyst patients [3-7], and extraction of pancreatic cyst information from narrative reports using the REX (Regenstrief EXtraction Tool) [8]. In this paper we modified the cyst identification technique using Unstructured Information Management Architecture (UIMA) pipeline and extended the cyst identification to feature extraction. REX uses a window of words before and after each concept to determine its negation status. We used NegEx algorithm [9], and improved it using Stanford dependency parser (SDP) [10].

Materials and Methods

Data Set:

Regenstrief Institute database, which contains the Indiana University (IU) health medical records, was queried for any text reports that contain the terms 'pancreas' and 'cyst' somewhere in the report. From the patients with 'pancreas' and 'cyst' in a text report, all of their text reports of any type were extracted. Then, 44 patients were selected at random forming a corpus with 1064 reports. The average number of reports per patient in our corpus is 23.64 with minimum of 1 and maximum of 221 report. Pancreatic cysts were present in 28 out of 44 patients and absent in the remaining 16 patients' records. The data set was randomly divided into two sets of train and test set, each containing 14 patients with pancreatic cyst and 8 patients with no pancreatic cysts. The test set is composed of 703 patient records, of which 509 belong to patients having pancreatic cysts and the remaining 194 records to patients without diagnosis of pancreatic cysts. The train set is composed of 316 patient records, of which 240 reports belong to patient having pancreatic cyst and the remaining 121 records to patients without pancreatic cysts.

Two pancreatic-cyst surgeon experts created the gold standard data. The discrepancies between the two annotators were resolved by discussing the differences in annotation. Cohen's kappa was used to measure the inter annotator agreement K=88% (11).

Methods:

UIMA was used as the framework for developing our text-mining system. UIMA is a software architecture developed by IBM for the analysis of unstructured content that includes text, audio or video data. It is openly available through the Apache software foundation [12]. UIMA is composed of processing units called analysis engine (AE) that analyze unstructured data and infers information from them. AEs are constructed from analysis logics called annotators. An AE may contain a single annotator (called primitive AE) or a combination of multiple primitive AEs and therefor contain multiple annotators (called aggregate AE).

Figure 1 shows the UIMA pipeline integrated to extract the information from medical reports. The input goes through a series of tasks depicted in each block. The input for every step is the output of its predecessor task.

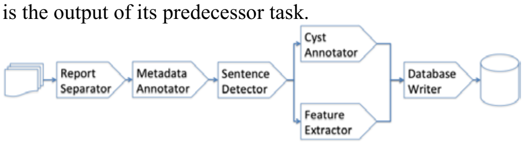

Figure 1: AEs used in the UIMA pipeline

The initial input is a text file containing all the reports for all the patients. Report separator AE in this pipeline separates each report. Each report has metadata information (medical record number (MRN), report id, report name, report date) and a report body which is the note dictated by physicians. Metadata annotator AE extracts these metadata from each report using regular expressions. Regular expression or regex is a pattern of characters matching specific strings of text. The next AE is the cTAKES sentence detector [13]. This is a UIMA wrapper around the openNLP sentence detector [14], which was originally used during the first pass. Because sentences were breaking on new lines, it was replaced with Ytex sentence detector. Ytex sentence detector is a modified version of cTAKES to deal with breaking sentences on new lines [15].

The cyst annotator AE extracts cyst concepts within each report using regular expression. Table 1 shows each concept and the regular expression developed to extract it from the patients' medical reports. Concepts that are used to identify pancreatic cysts in medical records were assembled by our medical team and additional keywords were added by searching through literature and Unified Medical Language System (UMLS) knowledge base [8]. The final sets of modifications were applied by analyzing the false positive and false negative results of the initial set of regular expressions that were applied to the training set. The same principles were applied in developing the feature extractor using regular expressions. Our medical team assembled an initial list of features important in the surveillance of patients with pancreatic cyst and predictive of pancreatic cancer. Further modifications were made based on review of literature and discussions on the required adjustments. Below is the list of features used to extract from the patient records.

Physician clinical notes:

High Risk Features:

- Symptoms/signs/conditions: pancreatitis, diabetes, steatorrhea, diarrhea, jaundice, weight loss
- Social history: Smoking, obesity
- Family history: pancreatic cancer
- Age >60

Low Risk Features:

- Absence of high risk features

Laboratory Reports:

High risk Features:

- Carcinogenic Antigen (CA) 19-9 >37 ng/ml
- Alkaline phosphatase >125 U/L
- Ha1c> 5.9 ng/ml
- Amylase >161 U/L
- Lipase >51 U/L
- Bilirubin >1 mg/dL

Low Risk Features:

- Absence of high risk features

Radiology Reports:

High risk Features:

- Mural Nodule(s)
- Thick Septa(e) $\geq$2mm
- Mass
- Main pancreatic duct dilation >5 mm

Low Risk Features

- Absence of high risk features

Pathology Reports:

High risk Features:

- Moderate or High grade Atypia
- KRas mutation
- Loss of heterozygosity

Low Risk Features

- Absence of high risk features

Table 1 – Various cyst types and their associated regular expression patterns

Concept	Regular Expression
Pancreatic Cyst	(?i)(pancreatic cyst(s)? | cyst(s)?(of| in)? the pancreas | pancreatic cystic)
Pancreatic pseudocyst	(?i)(pseudo\s?cyst(s)?)
Mucinous Cyst	(?i)(mucinous cyst(ic | ts) neoplasm | mucinous cystadenoma | intraductal papillary mucinous | \b(mcn)\b | \b(mca)\b | \b(ipmn)|\b(ipmt)\b)
Serous Cyst	(?i)(serous cyst(ic | s | adenoma) | \b(sca)\b)
Retention Cyst	(?i)(retention cyst(ic | s))
Cystic neuroendocrine tumor	(?i)(cystic neuroendocrine tumor | cystic neuroendorine | neuroendocrine cyst(ic|ts)|islet cell cyst tumor|cystic islet cell tumor)
Cystic degeneration cancer	(?i)(cystic degeneration cancer | cystic degeneration | degeneration cyst(ic | s))
Duct ectasia	(?i)(duct(al) ectasia | ectasia of the(pancreatic)? duct | ectasic duct)
Duct dilatation	(?i)(pancreatic duct(al) dilatation | dilatation of the(pancreatic)? duct | dilated(pancreatic)? duct)

Once the concepts and features were extracted, the NegEx algorithm was used to find out if they are negated or not. Negation detection was used as a method inside the cyst and feature annotator AEs.

The cyst annotator and feature extractor AEs are independent from each other and get the sentence and metadata related to concepts or features from proceeding AEs. The last annotator (Database Writer) receives all the information extracted in the previous sections (MRN, Report ID, Report Name, Report

Date, Sentence, Concept, Features, Negation) and writes them to a database.

Cyst Identification Scoring

There are multiple factors, which determine whether or not a patient has a diagnosis of pancreatic cyst. As mentioned in the introduction section, there are various types of pancreatic cysts but not all are considered malignant. Additionally, pancreatic cysts can appear in various reports once or multiple times. Finally the date of diagnosis can span between 1999 and 2012.

In order to account for all these elements, we have developed the following weighted cyst identification scoring based on pancreatic cyst type, report type, report date and frequency of the pancreatic cyst citation in every report.

$$\psi\,(C) = LOG_{10}\left(t(ct)^{r(c)} * \left[\frac{\#\ of\ c}{t(ct)}\right] * 100\right) * w(ct) \quad (1)$$

Where:

C: A cyst concept
ct: A cyst type in which the cyst concept is a member of
r(c): ranking factor for the cyst concept of c in the cyst type of *ct*.
# of c: total count for a cyst concept in the patient's records.
w(ct): a weighted factor value for the report types.
t(ct): total frequency count for the cyst in all reports of patient's longitudinal health record calculated using the following formula (If *N(ct)* is zero, *t(ct)* is considered one, to avoid infinity.)

$$t(ct) = \frac{P(ct)}{N(ct)} + w(t) \quad (2)$$

P(ct): number of cyst concepts that are affirmed.
N(ct): number of cyst concepts that are negated.
W(t): weighted factor for the report date.

The gold standard for pancreatic cyst identification is the pathology report followed by the radiology report including CT, MRI and EUS. These types of reports are favored followed by other types of reports such as clinical notes. A weighted factor was assigned to the report types based on their priority. A report conveying the date of the pancreatic cyst diagnosis is another important factor since our data extends over a period of 22 years. More recent diagnoses are favored higher in this formula. Finally, a higher term frequency of pancreatic cyst is considered if it has been mentioned often in the report.

NegEx and Stanford Dependency Parser

In order to improve the accuracy of negation algorithm, we used the SDP. The SDP finds out how words are related to each other in a sentence. Dependency is a binary asymmetric relation that holds between a token (word) and its dependents in a sentence.

The concept and the sentence containing it, are passed through the NegEx algorithm to identify the negation status of the concept. If the concept is negated, SDP was used to double-check the negation status. NegEx is a string-matching algorithm that looks for negation terms such as 'No', 'No evidence of', 'Rule out', etc., within the sentence containing the concept. Sometimes the negation terms do not refer to the concept in the sentence but the NegEx algorithm identifies the concept as negated. For instance, consider the sentence: *"The patient developed pseudocyst, which was drained by cyst gastrostomy and symptoms resolved."* Because of the negation term *"resolved"*, the NegEx algorithm will consider the identified concept *"pseudocyst"* as negated while the term *"resolved"* is referring to the *"symptoms"* not the *"pseudocyst"*.

Figure 1 shows the above sentence's dependency graph *produced by GrammerScope* [16]. In this sentence "pseudocyst" is the extracted concept and *"resolved"* is the negation term. As shown in figure 1, there is no relation between *"resolved"* and *"pseudocyst"*, therefore the sentence is not negated.

Usually there is no direct relation between the concept and negation terms in SDP. For instance, consider the sentence: *"The pancreatic survey demonstrates no evidence for pseudocystic or inflammatory changes."* When parsed through the SDP an output shown in figure 2 is generated. In this example, there are three levels of nested relations between the terms "*no evidence*" and "*pseudocystic*" (see table 2). We construct a production chain of the form *"no ➔ evidence ➔ changes ➔ pseudocyst"*. After manual analysis of the number of sentences, we found that the appropriate level to search for the relation between a concept and negation terms is three. The negation analysis is based on the output from the SDP. If we do not find the relation between the negation term and the concept within the three levels, we consider that concept affirmed, else it is considered negated.

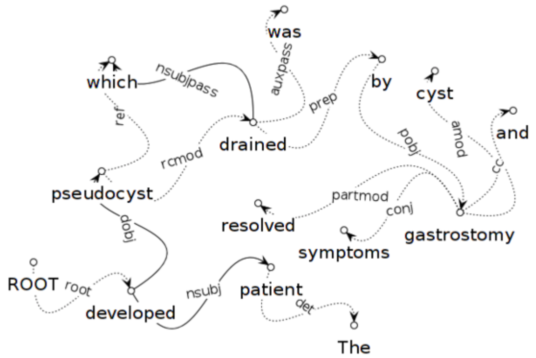

Figure 1- SDP for the sentence: "The patient developed pseudocyst, which was drained by cyst gastrostomy and symptoms resolved."

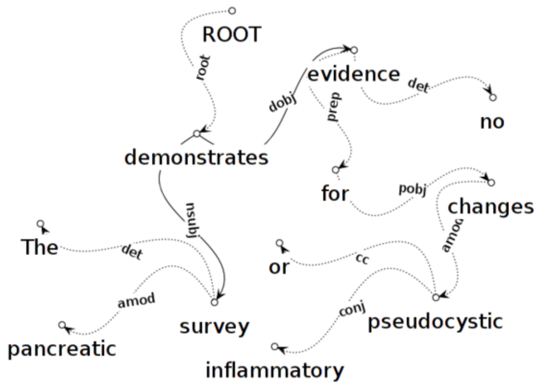

Figure2- SDP for the sentence: 'The pancreatic survey demonstrates no evidence for pseudocystic or inflammatory changes'

Exceptions in Handling SDP Dependencies

The production chain rule is not used in certain cases, for instance, when the negation term is the root term in the SDP tree, every dependent word at the root will be considered as negated. If the concept term is found dependent on the root (the negation term in this case), the concept is considered negated, otherwise it is affirmed. In another case where we consider the collapsed representation of SDP, if the preposition is a negation term, the dependent words in the relation are considered negated.

Table 2- Typed dependencies for the sentence: 'The pancreatic survey demonstrates no evidence for pseudocystic or inflammatory changes'

Chain order	Stanford Dependency	Token-Token index	Token-Token index
	det	survey-3	The-1
	amod	survey-3	pancreatic-2
	nsubj	demonstrates-4	survey-3
	root	ROOT-0	demonstrates-4
1 ->	det	evidence-6	no-5
	dobj	demonstrates-4	evidence-6
3 ->	amod	changes-11	pseudocystic-8
	cc	pseudocystic-8	or-9
	conj	pseudocystic-8	inflammatory-10
2 ->	prep_for	evidence-6	changes-11

Results

The regular expression was applied to the test set that is composed of 703 reports and the results were compared with the gold standard. Table 3 shows the number of affirmed (cyst exists) and negated (cyst does not exist) instances detected by the system and gold standard (manual annotation).

Table 3 –Comparison of gold standard and NLP system using NegEx algorithm

System Output	Gold Standard	
	Affirmed	Negated
Affirmed	TP=89	FP=1
Negated	FN=11	TN=10

Based on the true positive, true negative, false positive, and false negative scores shown in table 3, the precision, recall and F- Mesaure obtained are 98.9%, 89% and 93.6% respectively. The true negative rate is 90.9% and accuracy is 89.1%.
The results for SDP-NegEx based method for pancreatic cyst identification are given in table 4. The performance of SDP-NegEx significantly improved the performance compared to the NegEx algorithm alone. The number of false negatives dropped from 11 to 4 and therefore the recall, true negative rate, F-Measure, and accuracy has increased to 95.7%, 94.44%, 97.57% and 95.49% respectively. The precision stayed at 98.9% since the algorithm was not used if the concept was affirmed by NegEx algorithm.

Table 4 –Comparison of gold standard and NLP system using NegEx algorithm and SDP

System Output	Gold Standard	
	Affirmed	Negated
Affiremd	TP=89	FP=1
Negated	FN=4	TN=17

We repeated the same procedure for feature extraction with and without the improvement over the NegEx algorithm. Table 5 and 6 show the result of the system in comparison with the gold standard.

Table 5 –Comparison of gold standard and NLP system using NegEx algorithm

System Output	Gold Standard	
	Affirmed	Negated
Affirmed	TP=1519	FP=23
Negated	FN=40	TN=193

The precision and recall for the system with NegEx algorithm is 98.5% and 97.43% respectively. The F-measure is 97.96%, true negative rate is 89% and accuracy is 96.45%. The SDP-NegEx based method improved the results by decreasing the number of false negatives, from 40 to 29, and therefore increasing the number of true negatives from 193 to 204. The precision and recall based on table 6 are 98.5% and 98.12% respectively with an F-Measure of 98.31%. The accuracy is 97.07% and true negative rate is 90%.

Table 6 –Comparison of gold standard and NLP system using NegEx algorithm and SDP

System Output	Gold Standard	
	Affirmed	Negated
Affirmed	TP=1519	FP=23
Negated	FN=29	TN=204

Discussion

Pancreatic cancer is a deadly cancer due in part because it is typically diagnosed in advance stages when there is no effective treatment. Early detection of pancreatic cancer is possible through surveillance of patients at risk. Patients with pancreatic cysts or a hereditary predisposition to pancreatic cancer are considered to be at higher risk of developing cancer. In this retrospective study, we used NLP to find patients with pancreatic cysts. We used regular expressions with the Negex algorithm within a UIMA pipeline to identify cyst citation in various report types. We improved the negation detection using SDP using a chain of nested dependency relations between the Negex terms and cyst concepts or features. Dependency parser based negation (DepNeg) detection has been already used in clinical narratives but did not improve the performance of NegEx algorithm significantly [17]. Our system differs with DepNeg in that it double checks the negation if the NegEx algorithm, finds out that the concept is negated, and will not run if the concept is affirmed. DepNeg uses manually constructed rules to find out when concepts are negated. We use a chain of nested dependency relations between the concept and negation terms identified by the NegEx algorithm, which significantly improved our recall results.

We also identified and extracted a list of data elements that are important in the assessment of stratification of pancreatic cancer risk in patients with pancreatic cyst from their medical records. In order to detect the negation status of these features, NegEx algorithm and our improved negation algorithm was used, improving the recall from 97.43% to 98.12%.

Conclusion and Future Works

A new infrastructure has been created with a flexible architecture based on the UIMA platform. The goal is to identify patients with pancreatic cyst and risk stratify them into two groups of low and high risk of pancreatic cancer in the following modes of execution:

Surveillance
The processing pipeline will be used in a surveillance capacity. When collecting data in this manner, items such as notes and radiology reports could flow through the pipeline as they are added to the patient record in the Indiana Network for Patient Care (INPC) database. INPC was created by Investigators at the Regenstrief Institute in 1995 with the goal of providing clinical information at the point of care for the treatment of patients [18-19].. It is currently an operational community-wide secure data exchange.

Real-time

Real-time processing means that the NLP pipeline typically will process a small subset of documents within a limited scope (i.e., a single patient) to provide instant results. This mode of operation would typically be reserved for the Regenstrief physician order entry system to provide information to a clinician during the patient encounter.

Batch

In a batch setting, targeted filters would be employed to search over the existing clinical data within the patient data repository. This mode of operation is research focused to provide retrospective analysis that can be executed in specific points in time. The infrastructure is being built to provide a self-service approach to allow the configuration of the UIMA pipeline and results ontology by the researcher.

The initial testing of the UIMA pipeline will be done using the batch mode of operation. The first phase will be run using a cohort of patients to examine performance and accuracy. The second phase will be run on patients that have had clinical encounters for the day in the IU Health System and run on a daily basis. The objective for this phase would be to determine candidates for follow up care and to provide results to the specialists for review. The final goal would be to attain a degree of accuracy and performance to provide real time alerting to a patient's primary care provider during an encounter. The alert could be provided to the INPC via Docs4Docs (D4D) [20] or as an alert within a physician order entry system such as the Gopher currently in use at Wishard Hospital. D4D, which was implemented in 2004, is an innovative clinical messaging system that currently delivers patient clinical information to over 800 Indiana clinicians via web, fax or as HL7 messages directed to EMR. Replacing the traditional delivery of laboratory results such as courier or postal mail, D4D efficiently delivers aggregated clinical results to physicians regardless of where they practice through the INPC. Feedback from providers on D4D is standardized and will be employed. In addition, through D4D we will gather spontaneous feedback as well as formal feedback from providers regarding receipt, timeliness and accuracy of pancreatic cyst identification, stratification and screening/surveillance recommendation data.

Acknowledgments

This work was supported in part by the Agency for Healthcare Research and Quality R01 HS19818-01 and by a grant from the Office of the Vice President for Research at IUPUI.

References

[1] Gorin AD, Sackier JM. Incidental detection of cystic neoplasms of the pancreas. Md Med J. 1997: 46(2): 79-82.

[2] Siegel R, Naishadham D, Jemal A. Cancer statistics. CA Cancer J Clin. 2012: 62: 10–29.

[3] Cauley CE, Waters JA, Dumas RP, Meyer JE, Al-Haddad MA, DeWitt JM, Lillemoe KD, Schmidt CM.Outcomes of primary surveillance for intraductal papillary mucinous neoplasm. J Gastrointest Surg. 2012: 16(2): 258-67.

[4] Pandol S, Gukovskaya A, Edderkaoui M, Dawson D, Eibl G, Lugea A. Epidemiology, risk factors, and the promotion of pancreatic cancer: role of the stellate cell. J Gastroenterol Hepatol. 2012: 27 (Suppl 2): 127–134.

[6] Al-Haddad A, Schmidt CM, Sandrasegran K, Dewitt. Diagnosis and Treatment of Cystic Pancreatic Tumors. Clin. Gastroenterol. Hepatol. 2011: 9: 635–48.

[7] Waters A, Schmidt. Intraductal Papillary Mucinous Neoplasm—When to Resect? Adv Surg 2008: 42:87–108.

[8] Al-Haddad MA, Friedlin J, Kesterson J, Waters JA, Aguilar-Saavedra JR, Schmidt CM. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. HPG (Oxford). 2010: 12(10): 688-95.

[9] Chapman W, Bridewell , Hanbury , Cooper GF, Buchanan G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inform 2001: 34 (5): 301–10.

[10]de Marneffe MC, MacCartney, Manning D. Generating Typed Dependency Parses from Phrase Structure Parses. LREC. 2006.

[11]Carletta J. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics. 1996; 22(2).

[12]Apache Foundation The Unstructured Information Management Architecture. [Online]. [cited 2012 November 24]. Available from: http://uima.apache.org.

[13]Savova K, Masanz J, Ogren , Zheng , Sohn , Kipper-Schuler C, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010: 17(5): 507-13.

[14]OpenNLP. [Online]. [cited 2012 Nov 24]. Available from: http://opennlp.sourceforge.net/index.html.

[15]Garla Lo Re III , Dorey-Stein, Kidwai, Scotch M, Womack J, Justice A, Brandt C. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc. 2011: 18(5): 614-20.

[16]Bou. GrammerScope. [Online]. [cited 2012 December 6]. Available from: http://grammarscope.sourceforge.net.

[17]Sunghwan S, Stephen W, Chute G. Dependency Parser-based Negation Detection in Clinical Narratives. In AMIA Summits Transl Sci Proc 2012: 1–8.

[18]Biondich PG, Grannis SJ. The Indiana network for patient care: an integrated clinical information system informed by over thirty years of experience. J Public Health Manag Pract 2004: S81- 6.

[19]McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, Mamlin B, INPC Management Committee. . The Indiana network for patient care: a working local health information infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. Health Aff (Millwood). 2005: 24(5): 1214-20.

[20]Barnes. Lessons Learned from the Implementation of Clinical Messaging Systems. In AMIA Annu Symp Proc 2007: 36–40.

Address for correspondence

School of Informatics, Indiana University 719 Indiana Ave. Indianapolis, IN 46202, Tel: 317-278-7689
Email address: mpalakal@iupui.edu (Mathew Palakal)